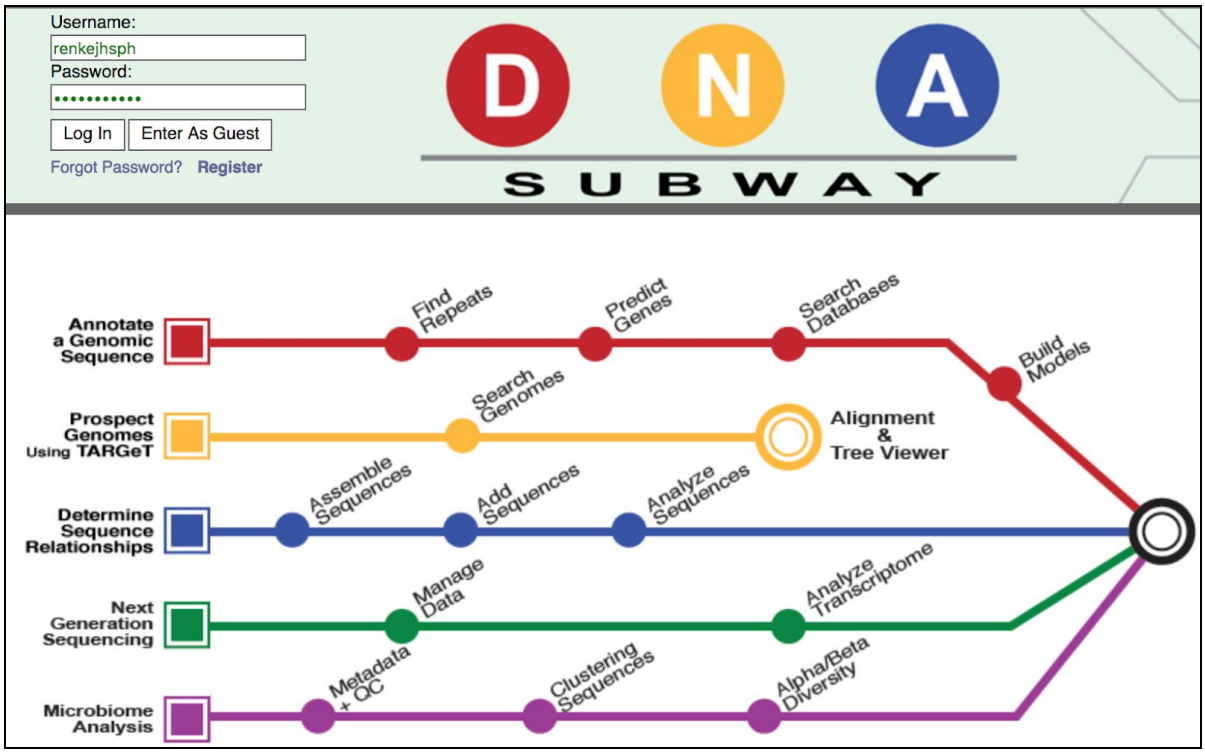
December, 2018

DNA Subway Purple Line Metagenome Analysis

Ray A Enke

Purple Line Metagenome Final Project
Dr. Enke Bio 481/581 Genomics

DNA Subway is collection of open access user friendly bioinformatics tools for various types of DNA sequence analyses. To learn more about the different types of tools available through DNA Subway, view this short ~2 min introductory video introducing the DNA Subway suite of bioinformatics tools.



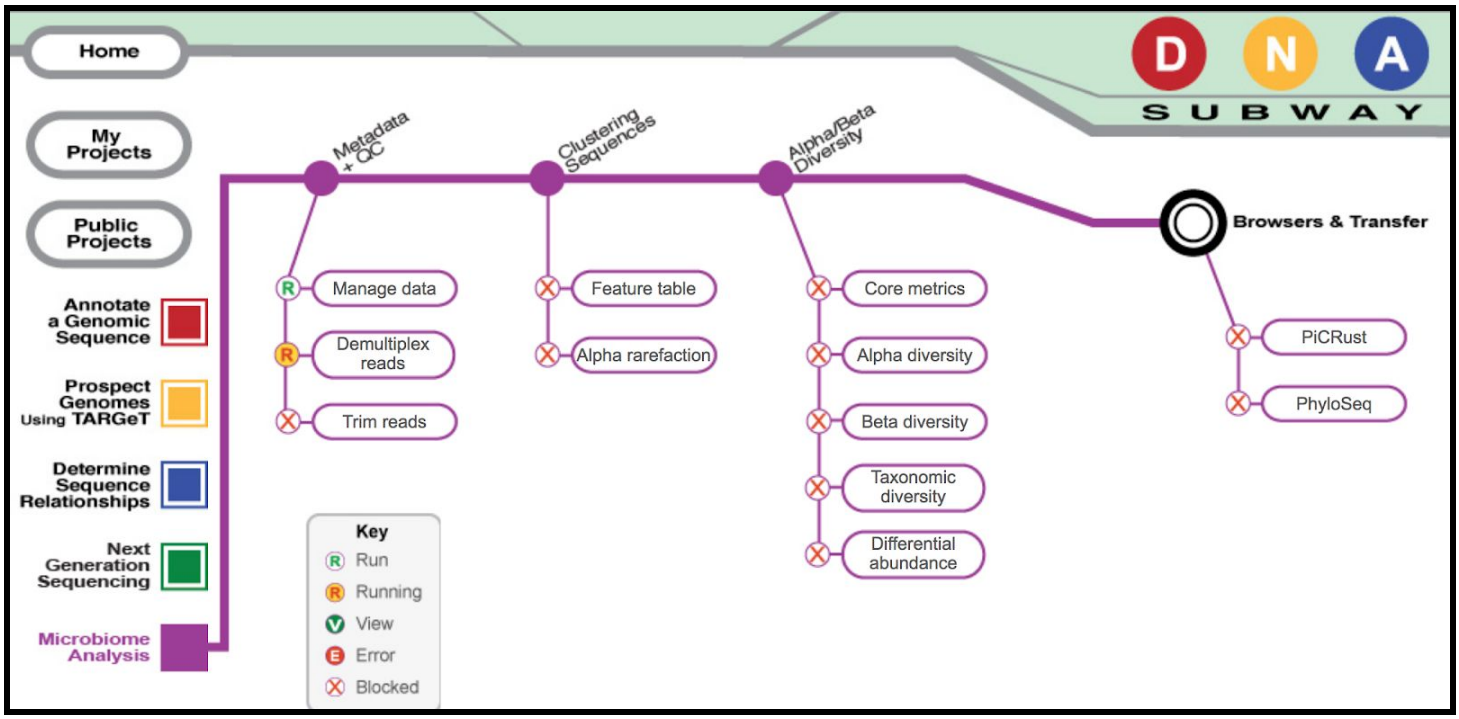
The **Purple Line** is a recent addition to DNA Subway that provides capability for **metagenomics** analysis of **microbiomes**. The Purple Line implements a simplified version of the QIIME 2 (pronounced “chime two”) workflow. Using the Purple Line, you will upload and analyze Illumina sequencing reads to identify species in microbial DNA samples, a process called **metabarcoding** analysis. Metabarcoding uses Illumina sequencing to analyze hundreds of thousands of DNA barcodes from complex mixtures of DNA.

In a typical experiment, like our snake experiment, DNA is isolated from sterile swabs or material taken from different environmental locations or conditions. PCR is used to amplify a variable genomics region, such as the 16S ribosomal RNA genes. Illumina sequencing is then used to generate ~3-5 thousands sequencing reads/sample. We have completed all of these steps and are now ready to push our data through the Purple Line bioinformatics pipeline.

Working in pairs or groups of 3 create a new Purple Line project

- Log in to DNA Subway>select the purple “Microbiome Analysis” option to open a new project
- Select “single end reads” for the project type and “Illumina Casava 1.8 for the format
- Name your project “4 sample snake test + your group’s initials”

The Purple Line has 3 main stops 1) **Metadata + QC**, 2) **Clustering Sequences**, and 3) **Alpha/Beta Diversity**.



The **Metadata + QC** steps will allow us to select the files we want to analyze from our CyVerse DE data store, link them with a metadata file describing what the samples are, and analyze sequence read quality and trim the reads. Here's a link to an example metadata file: https://bit.ly/2sg93Bl

- Select "Manage Data"
- In the "Shared with me" folder navigate to the folder that I shared with you containing 4 FASTQ files and 1 .tsv metadate folder
    - renkejhsph>Fall18_Purple_Line>Enke16S_fall18_snake_microbiome>Fall18_4-sample_analyasis
- Select the 4 FASTQ files corresponding to 2 female cloaca samples and 2 male oral cavity samples for the "Input files"
- Select the "Enke_4_sample_metadata.tsv" file
- Select "validate sample mapping file" to begin the process of linking the samples to the metadata
- Then select "validate"

- Once this is done you should see something like this indicating that you're ready to proceed:

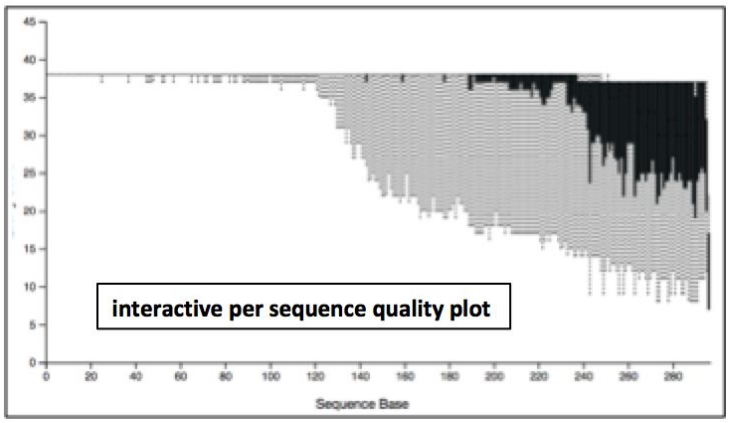
Input files:

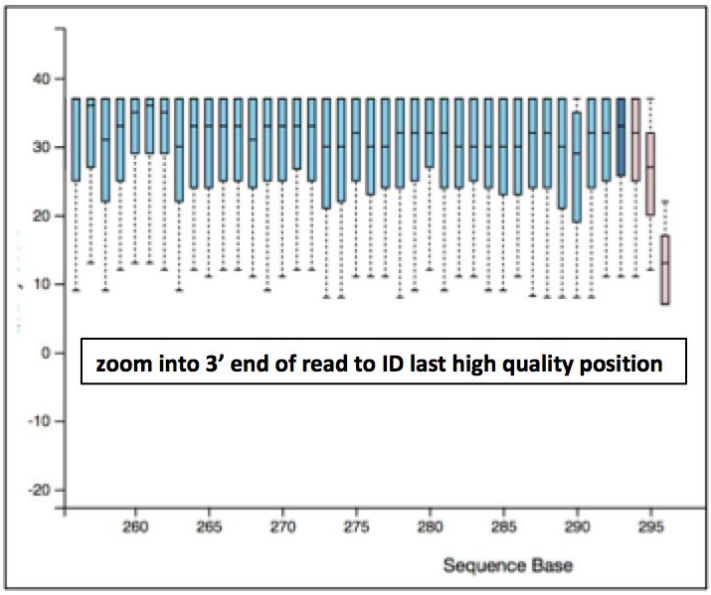
- [file] C2_sample_L001_R1_001.fastq.gz
- [file] D2_sample_L001_R1_001.fastq.gz
- [file] E6_sample_L001_R1_001.fastq.gz
- [file] F6_sample_L001_R1_001.fastq.gz
- [metadata] Enke_4_sample_metadata.tsv

add data
Validation done! :)

Next, you will **demultiplex** the samples. This is the process of taking the combined pool of sequencing reads and assigning them to specific samples by virtue of the sample-specific primer barcodes that were used during the PCR step of the library preparation.

- Click the 'Demultiplex reads' stop, then click demux reads to demultiplex your sample reads
- Keep all default settings and start the analysis (takes ~10 minutes)
- When demultiplexing is complete, you will generate an **imported-demux.qzv** file; click this link to view visualization and metrics on the sequence and metadata for this project
- **Collect the following data for your samples (needed for later steps in Purple Line)**
    - from the "Overview" tab, note the minimum & maximum # of "Demultiplexed Sequence Counts"
    - from the "Overview" tab, save a copy of the "Per-sample sequence counts table"
    - from the "Interactive Quality Plot" tab, click and drag on the interactive per sequence plot to view a close up of the last 30-40 bases on the 3' end of the read
    - For our upcoming trimming step, identify the base position where you see a shift in Phred quality score (from blue to red); note the last blue nt position for the trim step
    - save a copy of the full and zoomed in Per-sequence quality plots for your data

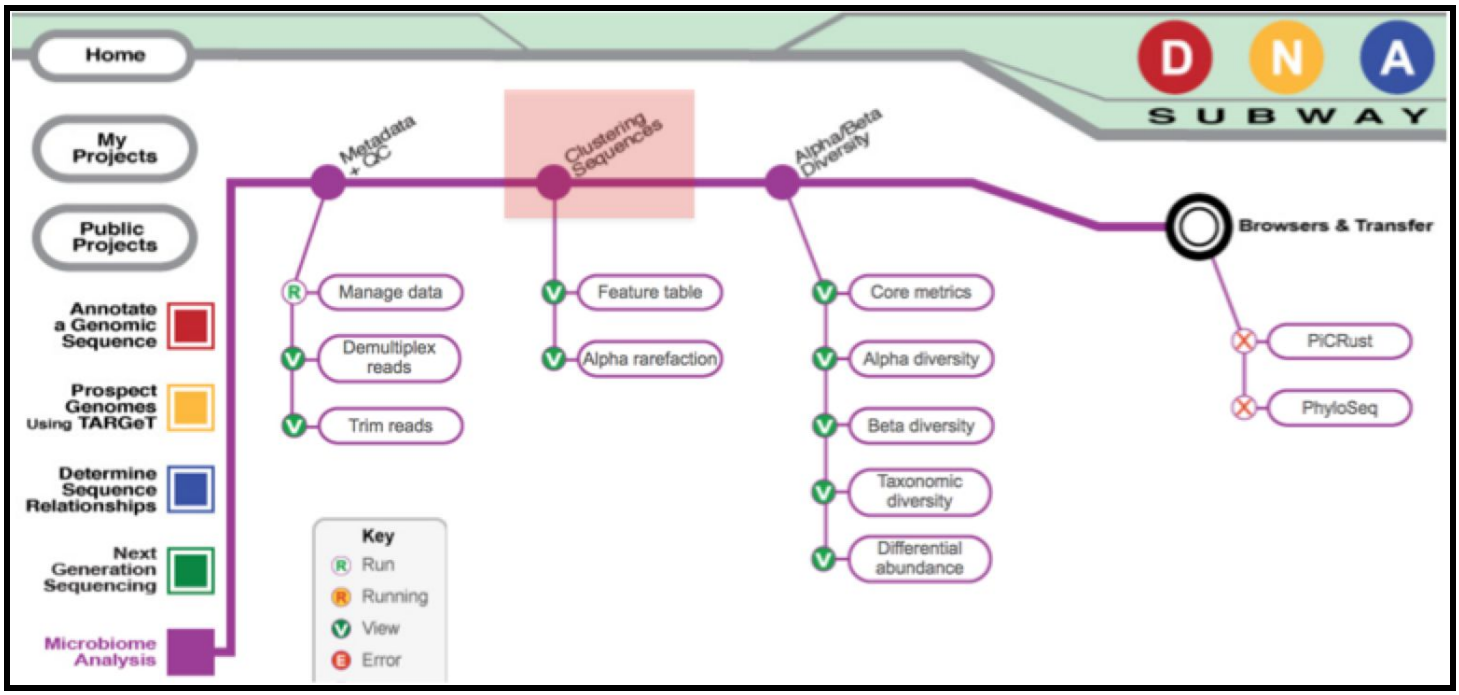




Next, click into the **Trim Reads** stop**.** This stop will filter out any low quality reads from our FASTQ data files from the 5' and 3' ends. The trim setting will depend on the data you collected from the demultiplex step.

- Click "run" and then select values for "trimLeft" (5' end) and "TruncLen" (3')
    - Set "trimLeft to 1
    - Set "TruncLen" to the base position where you saw a shift in Phred quality score on the 3' end in the Demultiplexing step
- Finally, click the "trim reads" link (will take a few minutes to run)
- When trimming is complete, you will generate a **table-trim.qzv** file; click this link to view
- From the "Overview" tab, you need 2 pieces of data from the **Frequency per sample** table to run subsequent Purple Line steps
    - **Minimum frequency =**
    - **Maximum frequency =**
    - These are the highest and lowest # of reads/sample in our 4 samples

Once your reads are trimmed, move onto the next set of Purple Line stops under the **Cluster Sequences** Heading.



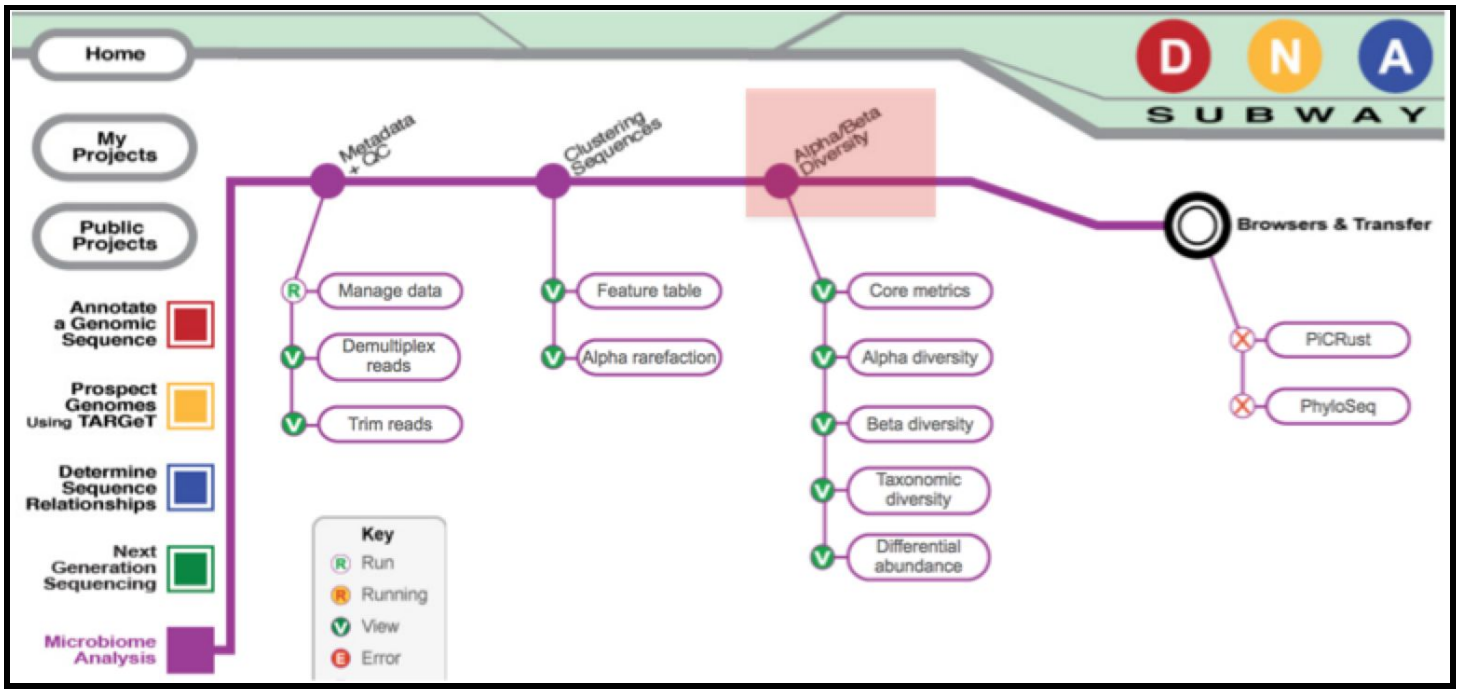
- Click **Feature table** and then the "Build feature table" link

The **Feature table** stop generates a list of all unique 16S sequences associated with your samples. When processed, you will get a link **rep-seqs.qzv file** visualizing each of these sequences. There are also links to download FASTA files of each sequences. We will not use these features for now.

Next click into the **Alpha Rarefaction** stop, this will tell Purple Line what range of sequencing reads we have per sample

- Select “run” and designate the **minimum** and **maximum rarefaction depth**
- A minimum value should be set at 1
- The maximum value is specific to your data set
    - this is the **Maximum frequency** metric you collected in the Trim step
- Click “submit job”

Once processed, the resulting **alpha-rarefaction-trim.qzv** plots demonstrate the #of reads per sample. We will ignore these plots for now and move onto the next set of sub steps under the **Alpha/Beta Diversity** step.



1. Click into **Core metrics** stop and select “run”
2. Set the **sampling depth**
    a. this is the **Minimum frequency** metric you collected in the Trim step
3. For the **Classifier** field, choose **Grenegenes (16s rRNA)** for microbiome analysis then submit

Once Core metrics analysis is complete, the next 3 stops in Purple Line are the payout of interesting data. These stops are:

4. **Alpha Diversity** - plots the overall diversity within each sample group (ie variation between sample replicates)
5. **Beta Diversity** - plots the overall diversity between each sample (ie variation between sample replicates and experimental groups)
6. **Taxonomic Diversity** - plots the specific microbial taxa associated with each sample

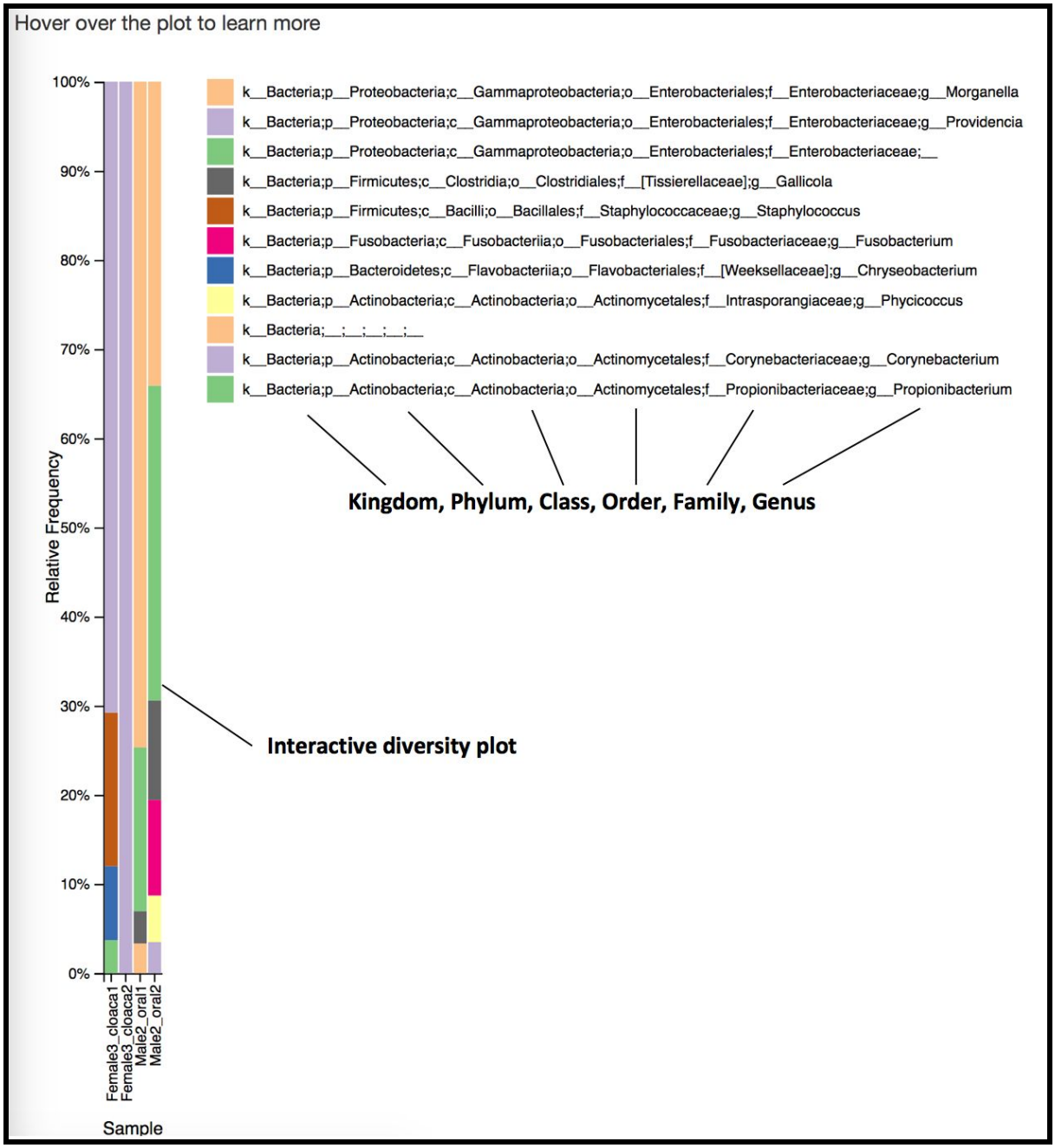
For today, we will focus on the **Taxonomic Diversity** stop which gives a readout of all microbes associated with our individual samples based on their 16S sequencing reads.

- Open the **Taxonomic Diversity** stop
- There are 2 .qvz links side by side, select the left **taxa-bar-plots.qzv** link

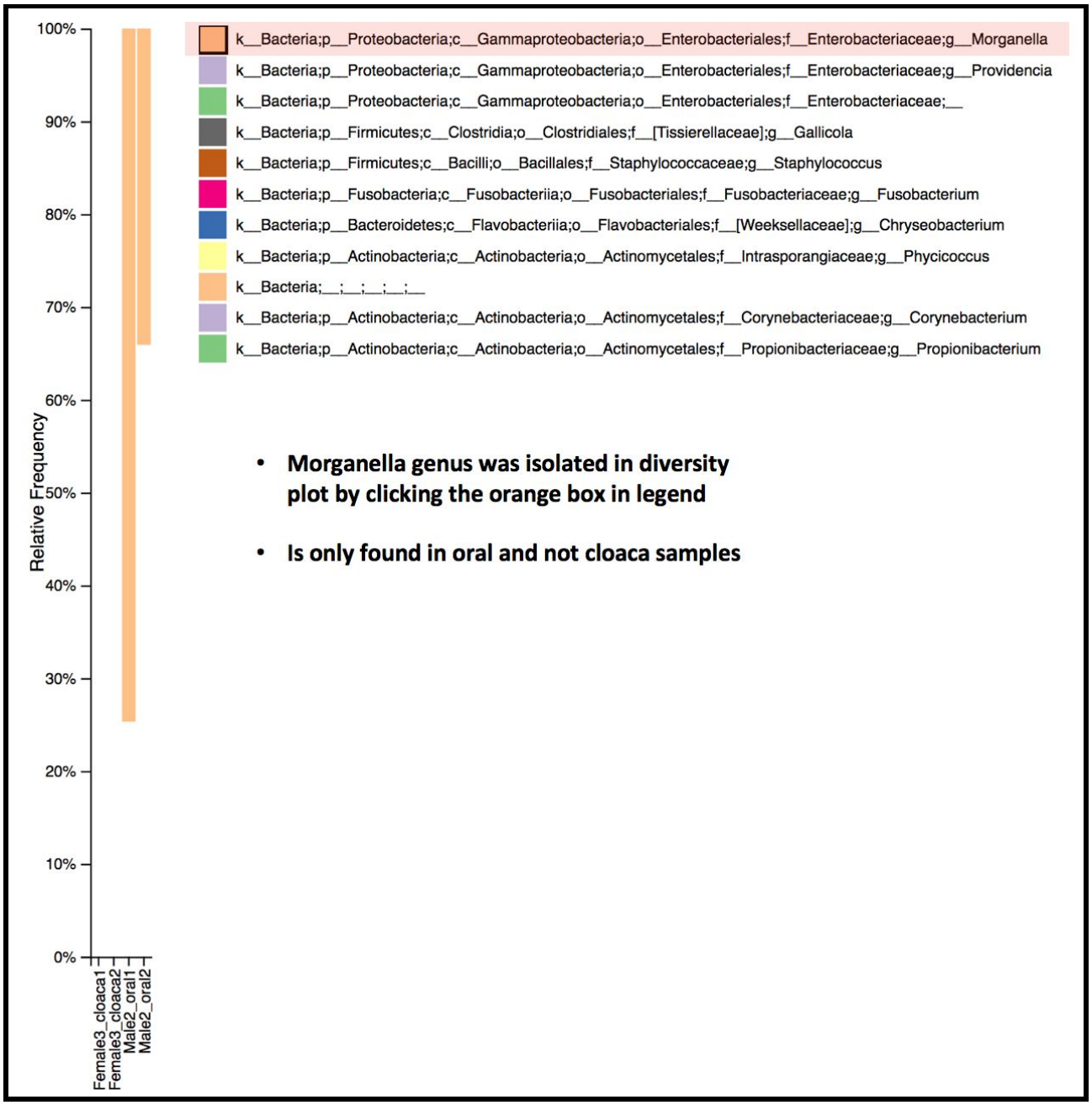
The **Taxonomic Level** option allows us to select which levels to visualize diversity in each sample (e.g. Kingdom, Phylum, Class, Order, Family, Genus, Species)

  - Select **Level 6** for the taxonomic level to visualize diversity at the **Genus level**

- For the **Sort Samples By** option, select **Description** to view diversity sorted by each sample



The resulting diversity plot lists all microbial taxa found in each sample. The taxa are color coded and listed in the legend on the right. The plot and the legend are both interactive. Hover over the bars in the plot to get a readout of the sample and individual microbial taxon. Click on individual taxon in the legend to isolate them in the plot.



- Which single genus of bacteria is unique to cloaca samples (found in both cloaca samples but neither oral sample)?

- Which two genera of bacteria are unique to oral cavity samples (found in both oral samples but neither cloaca sample)?